Discovery of a high-altitude ecotype and ancient lineage of Arabidopsis thaliana from Tibet

Journal:	*Science Bulletin*
Manuscript ID	CSB-2017-0708
Manuscript Type:	Short Communication
Date Submitted by the Author:	26-Jul-2017
Complete List of Authors:	Zeng, Liyan; Fudan University, School of Life Science Gu, Zhuoya; Fudan University Xu, Min; Tibet University Zhao, Ning; Tibet University Yonezawa, Takahiro; Fudan University Zhu, Weidong; Tibet University Liu, Tianmeng; Tibet University Zhang, Yang; university of Illinois at Urbana-Champaign Qiong, Lha; Tibet University Tersing, Tashi; Tibet University Xu, Lingli; Fudan University, School of Life Science Xu, Rongyan; Fudan University Sun, Ningyu; Fudan University Huang, Yanyan; Fudan University Lei, Jiankun; Fudan University Zhang, Liang; Fudan University Xie, Feng; Soochow University Zhang, Fang; 中科院遗传与发育生物学研究所

	Gu, Hongya; 北京大学生命科学学院 Geng, Yu peng; Yunnan University Hasegawa, Masami; Fudan University, School of Life Sciences Yang, Ziheng; University College London Crabbe, M. James C.; University of Oxford Chen, Fan; 中科院遗传与发育生物学研究所 Zhong, Yang; 复旦大学生命科学院,
Keywords:	Arabidopsis thaliana, divergence time, Qinghai-Tibet Plateau
Speciality:	Life & Medical Sciences/Plant Sciences

Discovery of a high-altitude ecotype and ancient lineage of *Arabidopsis thaliana* from Tibet

Liyan Zeng[1,2,3]†, Zhuoya Gu[2]†, Min Xu[1,4]†, Ning Zhao[1]†, Weidong Zhu[1], Takahiro Yonezawa[2,5], Tianmeng Liu[1], Lha Qiong[1], Tashi Tersing[1,6], Lingli Xu[2], Yang Zhang[7], Rongyan Xu[2], Ningyu Sun[2], Yanyan Huang[2], Jiankun Lei[8], Liang Zhang[8], Feng Xie[9], Fang Zhang[10], Hongya Gu[11], Yupeng Geng[12], Masami Hasegawa[2,5], Ziheng Yang[13], M. James C. Crabbe[14,15], Fan Chen[10,*], Yang Zhong[1,2,16,*]

1. Institute of Biodiversity Science and Geobiology, College of Sciences, Tibet University, Lhasa 850000, China

2. Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai 200433, China

3. Center for Clinical Public Health, Fudan University, Shanghai 201508, China

4. Institute of Forest Inventory, Planning and Research of Tibet Autonomous Region, Lhasa 850010, China

5. Institute of Mathematical Statistics, Midori-cho 10-3, Tachikawa, Tokyo 190-8562, Japan

6. Tibet Museum of Natural Science, Lhasa 850000, China

7. Department of Bioengineering, University of Illinois at Urbana-Champaign, Champaign, IL, 61801, USA

8. School of Computer Science, Fudan University, Shanghai 200433, China

9. School of Urban Rail Transportation, Soochow University, Suzhou 215131, China

10. Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

11. School of Life Sciences, Peking University, Beijing 100871, China

12. School of Life Sciences, Yunnan University, Kunming 650091, China

13. Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, United Kingdom

14. Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford, OX1 3PS, United Kingdom

15. Institute of Biomedical and Environmental Science & Technology, Department of Life Sciences, University of Bedfordshire, Park Square, Luton, LU1 3JU, United Kingdom
16. Shanghai Center for Bioinformation Technology, Shanghai 201203, China

†These authors contributed equally to this work
*Correspondence to: Fan Chen & Yang Zhong

Arabidopsis thaliana has long been a model species for dicotyledon study, and was the first flowering plant to get its genome completed sequenced [1]. Although most wild *A. thaliana* are collected in Europe, several studies have found a rapid *A. thaliana* west-east expansion from Central Asia [2]. The Qinghai-Tibet Plateau (QTP) is close to Central Asia and known for its high altitude, unique environments and biodiversity[3]. However, no wild-type *A. thaliana* had been either discovered or sequenced from QTP. Studies on the *A. thaliana* populations collected under 2000 m asl have shown that the adaptive variations associated with climate and altitudinal gradients [4]. Hence a high-altitude *A. thaliana* provides a precious natural material to investigate the evolution and adaptation process.

Here, we present the genome of a new ecotype of *A. thaliana* collected in the Gongga County, Tibet (4200 m asl) (Fig. 1A), to demonstrate its evolutionary history and adaptation to high-altitude regions. The Tibetan samples were identified as *A. thaliana* by comparing the nuclear internal transcribed spacer (ITS), four chloroplast genes (*mat*K, *rbc*L, *rpo*B, and *rps*16), and three chloroplast intergenic spacers (IGS, *trn*L-*trn*F, and *trn*T-*trn*L) with *A. thaliana* (Col-0) and *A. lyrata* (Supplementary Fig. 1). This is the first report that an *A. thaliana* population has been collected in the QTP over 4000 m asl and identified by molecular analysis. Moreover, the new Tibetan ecotype (herein referred to as "Tibet-0") is diploid (2n=10) according to karyotype analysis of its pollen mother cells during meiosis (Supplementary Fig. 2), suggesting that the ploidy of the Tibet-0 is stable and capable of further sequence analysis.

We then conducted genome-wide resequencing of Tibet-0 with a mean coverage of 40x of the reference genomes Col-0 and TAIR10, by using Illumina Hiseq2000 (Supplementary Table 5-6). We compared Tibet-0 with 47 other *A. thaliana* ecotypes that have been genome-wide sequenced, and found that Tibet-0 was of high divergence, including a higher proportion of SNPs (Supplementary Table 7-9). Evolutionary relationships between Tibet-0 and other ecotypes were evaluated by the following two independent approaches based on 5611 single-copy orthologues in 47 *A. thaliana* ecotypes including 26 relicts and 21 non-relicts defined by the 1001 Genomes Consortium[5]. The first approach is the phylogenetic method. The genealogy among the individuals were inferred based on the concatenated genomic data, and Tibet-0 was placed at the root of the *A. thaliana* populations with high support value (Fig. 1B) [5, 6]. It makes Tibet-0 the most ancestral lineage.

However, since this phylogenetic approach assumes that all gene loci have the same genealogy, coalescent method was also applied as a cross check [7]. In this method, 2788 single-copy orthologues were independently analyzed, and the distributions of the tMRCAs (the time to the most recent common ancestor) for these genes were estimated. If Tibet-0 is the most basal lineage among the *A. thaliana* populations and Tibet-0 specific alleles has generally older histories than others, tMRCAs excluding Tibet-0 will be smaller than tMRCAs of all *A. thaliana* populations. Otherwise, if there is no such genetic structure and Tibet-0 specific alleles are included within the genetic diversity of other *A. thaliana* populations,

tMRCAs excluding Tibet-0 will be equal to the tMRCAs of all *A. thaliana* populations. To examine the differences among the distributions, the tMRCAs were first estimated based on 48 *A. thaliana* (tMRCA48). Subsequently, each ecotype was excluded once, and the tMRCAs of 47 remaining *A. thaliana* were estimated (tMRCA47: there are 48 combinations of tMRCA47). Finally, the relative tMRCAs (tMRCA47/tMRCA48) were estimated. Fig. 1C illustrates the distributions of the relative tMRCAs. When Tibet-0 was excluded, the distribution of the relative tMRCAs ($tMRCA47_{excluding\ Tibet\text{-}0}$/tMRCA48) significantly shifted (t test, p = 9.77E-32), while the average of $tMRCA47_{excluding\ one\ ecotype\ other\ than\ Tibet\text{-}0}$/tMRCA48 showed no significant change (Fig. 1C). These findings confirm that Tibet-0 has the most ancestral positions among *A. thaliana* populations.

To understand the correlation between the evolution of *A. thaliana* and major geological events, especially Tibetan uplifts, the divergence time between Tibet-0 and other ecotypes were estimated. Since there is no suitable fossil calibrations within *A. thaliana*, the divergence time between *A. lyrata* and *A. thaliana* was estimated based on the genomic data in the framework of whole land plant evolution with reliable fossil records, and it was estimated to be about 9 million years ago (Fig. 1D, Supplementary Fig. 3). Then, the time of the common ancestor of *A. thaliana* was estimated by multiplying the divergence time between *A. lyrata* and *A. thaliana* and the ratio of the divergence time between *A. lyrata* and *A. thaliana*. The divergence time between Tibet-0 and other ecotypes was found to be 126 – 149 Ka (kili annum: thousand years ago). Interestingly, the Gonghe movement, which was the last phase of Tibetan uplift, isolated the Qinghai Lake and raised the QTP to its present height algo began at about 15 Ka [8]. Besides, the divergence time of Tibet-0 and other ecotypes is in the middle Pleistocene from 781 to 126 Ka [9].

A. thaliana has been widely used in studies of plant biology. By collecting and sequencing *A. thaliana* collected from the QTP over 4200 m asl, we have found that the Tibet-0 is a new and divergent ecotype that isolated from other *A. thaliana* ecotypes since the last uplift of the QTP. After 126 – 149 thousands years evolution in the extreme plateau environment, Tibet-0 possesses a distinctive genome with a high proportion of SNPs compared to other ecotypes. According to the strongly negatively skewed Tajima's D of 5611 single-copy orthologues, a recent selective sweep or population expansion might have occurred in the *A. thaliana*, which is consistent with previous studies (Supplementary Fig. 6)[10]. Considering the ancestral position of Tibetan populations as well as the subsequent selective sweep or population expansion, possibly in the Last Glacial Period, suggested by the negative Tajima's D, we suppose that some mutations might have emerged in the ancient *A. thaliana* population located around the QTP, and then spread to most other populations. Following step is investigating phenotypic traits of Tibet-0 to study the adaptive evolution of *A. thaliana* to high altitudes. As a new model plant, the Tibet-0 from QTP would provide an invaluable material for further study.

Declarations

Acknowledgements

This study was supported by the National Natural Science Foundation of China (91131901), the specimen platform of China (teaching specimens sub-platform) and PSCIRT project.

Availability of data and materials

The genomic DNA of Tibet-0 have been deposited in the Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra/) under accession number SRP052218.

Author Contributions

F. C. and Y. Z. conceived the project. L. Zeng, Z. G., T. Y., F. C. and Y. Z. contributed to the design of the project and extensive discussions. M. X., N. Z., W. Z., L. Q. and T. T. collected samples from Tibet. L. Zeng and H. G. helped with sample identification. L. X., R., X., F. X., J. L., L. Z., Z. G., N. Z., Y. H., T. Y., M. H., F. Z., F. C., Y. G., L. Zhang, Z. Y., M. J. C. C. and Y. Z. performed the common garden experiments, sequence analyses and evolutionary analyses. L. Zeng, Z. G., M. J. C. C., N. S., F. C. and Y. Z. wrote the manuscript. Other authors revised the manuscript.

Correspondence and requests for the materials should be addressed to F.C. (fchen@genetics.ac.cn) or Y. Z. (yangzhong@fudan.edu.cn).

Competing interests

The authors declare no competing financial interests.

Ethics approval and consent to participate

Not applicable

Supplementary Materials:

Figures S1-S4

Table S1-S9

Reference:

1. Arabidopsis Genome, I., *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* Nature, 2000. **408**(6814): p. 796-815.
2. Yin, P., et al., *The origin of populations of Arabidopsis thaliana in China, based on the chloroplast DNA sequences.* BMC Plant Biol, 2010. **10**: p. 22.
3. Liu S, W.N., Duan K, Xiao C, Ding Y, *Recent progress of glaciological studies in China.* Journal of Geographical Sciences, 2004. **14**(4): p. 401-410.
4. Suter, L., et al., *Gene regulatory variation mediates flowering responses to vernalization along an altitudinal gradient in Arabidopsis.* Plant Physiol, 2014. **166**(4): p. 1928-42.
5. Consortium, G., *1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana.* Cell, 2016. **166**(2): p. 481-91.
6. Gan, X., et al., *Multiple reference genomes and transcriptomes for Arabidopsis thaliana.* Nature, 2011. **477**(7365): p. 419-23.
7. Nakagome, S., S. Mano, and M. Hasegawa, *Comment on "Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage".* Science, 2013. **339**(6127): p. 1522.
8. Li, J., *Late Cenozoic intensive uplift of Qinghai-Xizang Plateau and its impacts on environments in surrounding area.* Quaternary Science, 2001. **21**: p. 381-391.
9. Cohen KM, G.P., *Global chronostratigraphical correlation table for the last 2.7 million years.* Subcommission on Quaternary Stratigraphy, 2011. **31**(2): p. 243-247.
10. Shimizu, K.K., et al., *Darwinian selection on a selfing locus.* Science, 2004. **306**(5704): p. 2081-4.

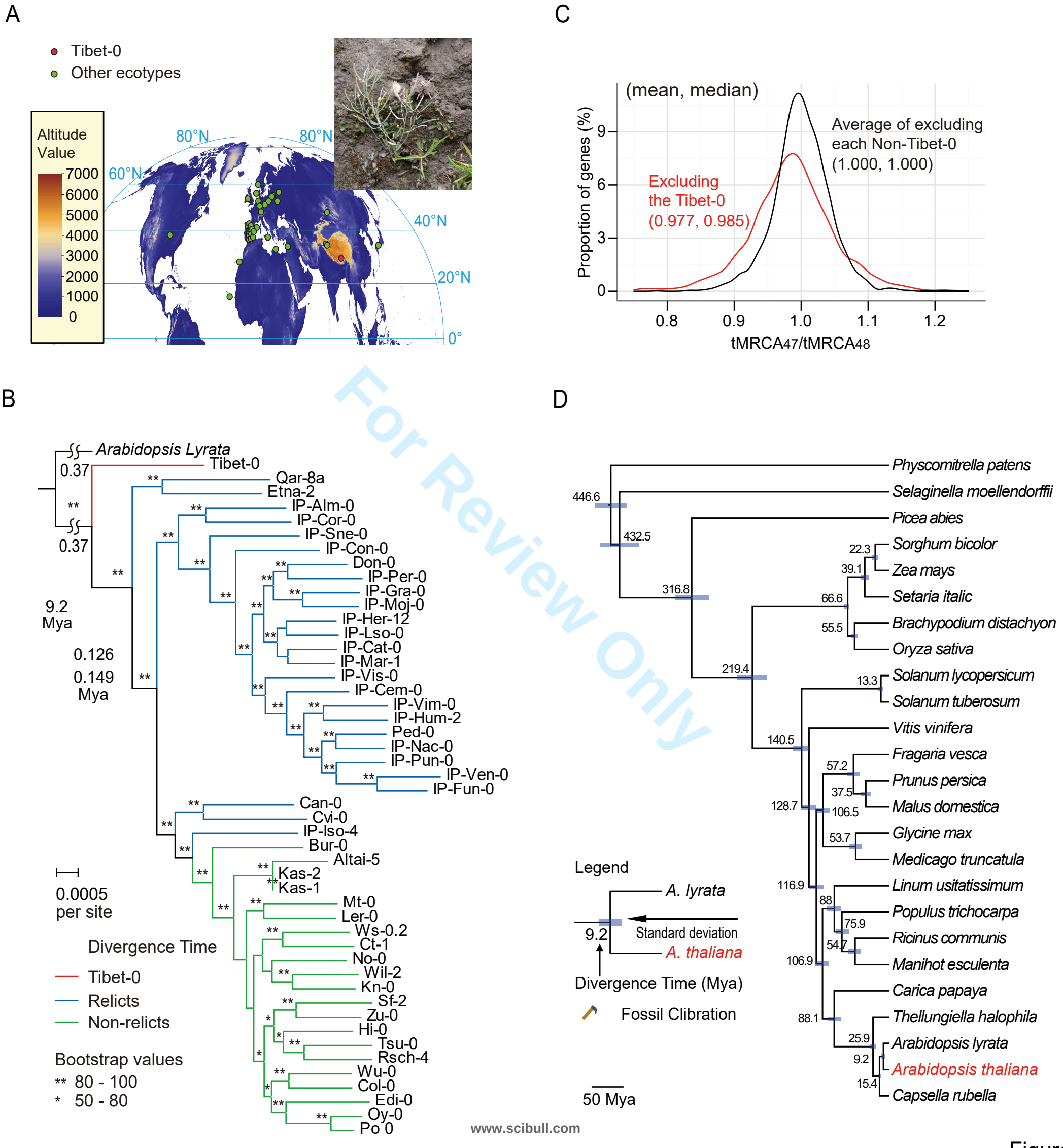


Figure1

Figure 1. Collection, phylogentic analysis and adaptation analysis of the Tibet-0 and other 47 *A. thaliana* ecotypes.

A, Origins of Tibet-0 and other 47 *A. thaliana* ecotypes that we have analyzed in this paper (Supplementary Table 4). Elevation data were downloaded from WorldClim (http://www.worldclim.org/). Colors indicate altitudes, going from low to high elevation: Deep blue, blue, white, yellow, orange, and brown.

B, The maximum likelihood phylogenetic trees based on 5611 orthologues of Tibet-0, 47 *A. thaliana* ecotypes and *A. lyrata* used as outgroup. Bootstrap values based on 100 replications are listed as percentages at each node. The Tibet-0 was marked in red.

C, The time of most recent common ancestor (tMRCA) based on 2788 single-copy orthologues. The red line represents the tMRCA47/tMRCA48 distribution where the tMRCA47 values exclude Tibet-0. The black line represents the mean of tMRCA47/tMRCA48 distribution where the tMRCA47 values exclude each Non-Tibet-0.

D, Phylogenetic affinities inferred from the maximum likelihood analysis of nucleotide sequence of 334 single copy orthologues in 25 plants. The divergence time of A. thaliana and A. lyrata was about 9.2Mya (million years ago). 7 fossil calibrations used in the study were marked as the hammer symbol. Branch lengths are proportional to the number of expected nucleotide substitutions. The number on the branch is the divergence time and unit is Mya.

A

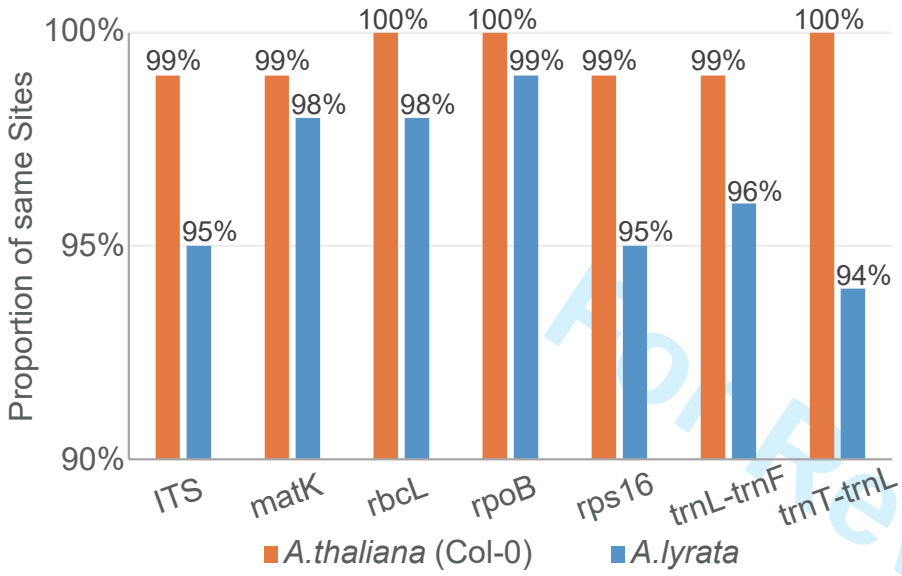
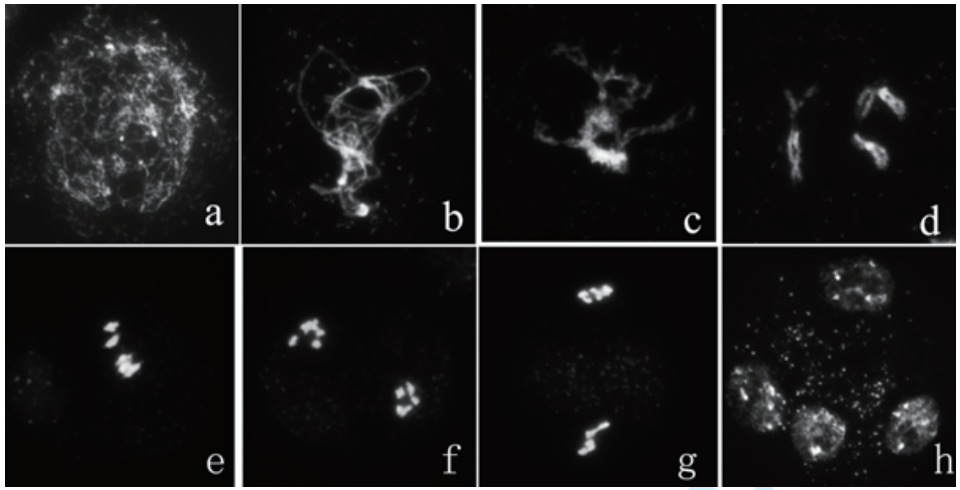


Figure S1

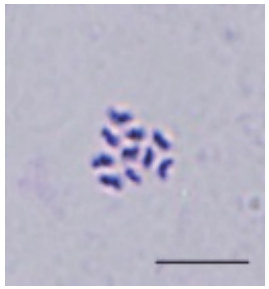
Figure S1. Gene barcoding comparison of Tibet-0 with *A. thaliana* and *A. lyrata*.
A. The barcoding genes were nuclear internal transcribed spacer (ITS), four chloroplast genes (matK, rbcL, rpoB, rps16) and two chloroplast intergenic spacers (IGS, trnL-trnF, trnT-trnL).

A

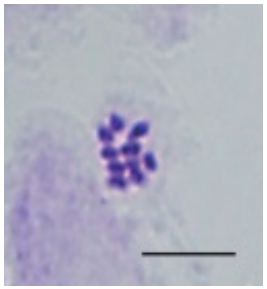


Tibet-0 n=2

B



Col n=2
Bar = 1um



Tibet-0 n=2

Figure S2

Figure S2. Results of the polyploidy analysis of the Tibet-0.

A. Each phase of the meiosis of Tibet-0 pollens was listed below. For each phase of Tibet-0, a: leptotene; b: pachytene; c: diplotene; d: diakinesis; e: Metaphase I; f: Anaphase I; g: Metaphase II; h: tetrads stage.

B. Karyotypes of *A. thaliana* ecotype Col-0 and Tibet-0.

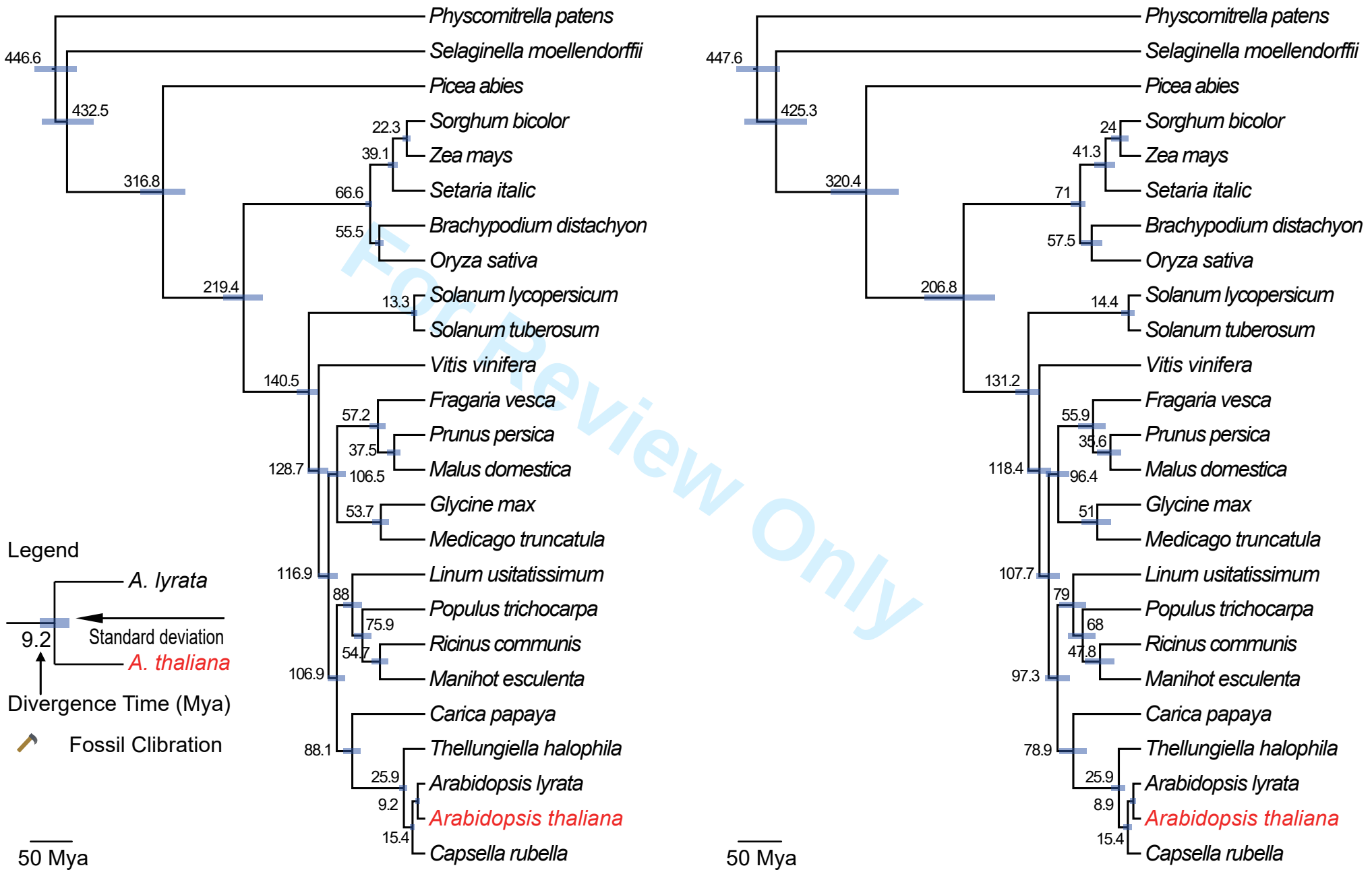


Figure S3

Figure S3. Results of divergence time estimation between *A. thaliana* and *A. lyrata*.
A. Phylogenetic affinities inferred from the maximum likelihood analysis of nucleotide sequence of 334 single copy orthologues in 25 plants. The divergence time of *A. thaliana* and *A. lyrata* was about 9.2Mya (million years ago). 7 fossil calibrations used in the study were marked as the hammer symbol. Branch lengths are proportional to the number of expected nucleotide substitutions. The number on the branch is the divergence time and unit is Mya.
B. Phylogenetic affinities inferred from the maximum likelihood analysis of amino acid sequence of 334 single copy orthologues in 25 plants. The divergence time of *A. thaliana* and *A. lyrata* was about 8.9Mya (million years ago). 7 fossil calibrations used in the study were marked as the hammer symbol. Branch lengths are proportional to the number of expected nucleotide substitutions. The number on the branch is the divergence time and unit is Mya.

A

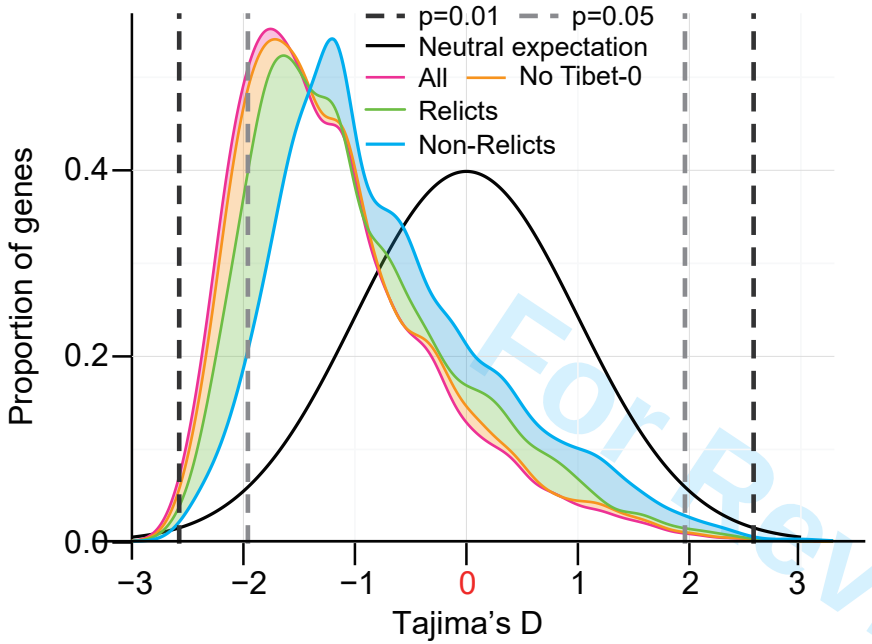


Figure S4

Figure S4. Results of the nucleotide diversity analysis of 48 *A. thaliana* ecotypes. A. Distributions of Tajima's D statistic for each of the 5611 genes of *A. thaliana* populations. The black line shows the expected distribution of D with no selection in a panmictic population of constant size. The red line shows the distribution of D of 5611 genes of Tibet-0 and other 47 *A. thaliana* accessions. The yellow line shows the distribution of D of 5611 genes of 47 *A. thaliana* accessions that excluding Tibet-0. The green line shows the distribution of D of 5611 genes of relicts. The blue line shows the distribution of D of 5611 genes of non-relicts that exclude relicts and Tibet-0.

Table S1. Amplification information of chloroplast and nuclear genes in Tibet-0 for barcoding

gene name	Primers sequence (5'-3')	Ta（℃）	size (bp)
ITS	TCCTCCGCTTATTGATATGC	55	682
	GGAAGTAAAAGTCGTAACAAGG		
matK	CGTACAGTACTTTTGTGTTTACGAG	55	908
	ACCCAGTCCATCTGGAAATCTTGGTTC		
rbcL	ATGTCACCACAAACAGAGACTAAAGC	50	710
	TCGCATGTACCTGCAGTAGC		
rpoB	AAGTGCATTGTTGGAACTGG	50	488
	GATCCCAGCATCACAATTCC		
rps16	GTGGTAGAAAGCAACGTGCGACTT	60	865
	TCGGGATCGAACATCAATTGCAAC		
trnL-trnF	CGA AAT CGG TAG ACG CTA CG	55	978
	ATT TGA ACT GGT GAC ACG AG		
trnT-trnL	CATTACAAATGCGATGCTCT	55	578
	TCTACCGATTTCGCCATATC		

Table S2. Twenty five species used in divergence time estimation

Species	Order_APGIII	Sources
Cucumissativus	Cucurbitales	Phytozome
Glycine max	Fabales	Phytozome
Medicagotruncatula	Fabales	Phytozome
Fragariavesca	Rosales	Phytozome
Malus domestica	Rosales	Phytozome
Pruspersicaun	Rosales	Phytozome
Manihotesculenta	Malpighiales	Phytozome
Ricinuscommunis	Malpighiales	Phytozome
Populustrichocarpa	Malpighiales*i*	Phytozome
Linumusitatissimum	*Malpighiales*	Phytozome
Arabidopsis thaliana	Brassicales	Phytozome
Arabidopsis lyrata	Brassicales	Phytozome
Capsella rubella	Brassicales	Phytozome
Carica papaya	Brassicales	Phytozome
Thellungiellahalophila	Brassicales	Phytozome
Vitisvinifera	Vitales	Phytozome
Solanum lycopersicum	Solanales	Phytozome
Solanum tuberosum	Solanales	Phytozome
Oryza sativa	Poales	Phytozome
Sorghum bicolor	Poales	Phytozome
Setariaitalica	Poales	Phytozome
Zea mays	Poales	Phytozome
Piceaabies	Gymnosperm	Spruce Genome Project
Selaginellamoellendorffii	Fern	Phytozome
Physcomitrella patens	Moss	Phytozome

Table S3. Seven fossil calibration used in divergence time estimation

Node/Fossil	Lower (most recent)	Upper (most ancient)	Calibration information
A/Cryptospore assemblage &Baragwanathialongifolia	421	472	Tims& Chambers1984; Garrat& Rickards 1987, Hueber 1992, Kenrick& Crane 1997, Rubinstein et al.2010, Magallón S et al. 2013
B/Perticaquadrifaria& P. varia	398		Gensel& Andrews 1984; Kenrick& Crane 1997, Magallón S et al. 2013
C/ Cordaitales	318		Phillips 1980, Taylor et al. 2009, Magallón S et al. 2013
D/Tricolpate pollen grains		130	Doyle et al. 1977, Hughes & McDougall 1990, Magallón S et al. 2013
E/Araceae&Zea/Oryza	65	84	Christopher 1979, Daghlian 1981, Piperno& Sues 2005, Prasad et al. 2005, Magallón S et al. 2013
F/Fagales&Fabales	85		Zhang et al. 2012
G/Sapindales	65		Zhang et al. 2012

Table S4. The location and altitude of Tibet-0 and 47 other ecotypes

Sample	country	Latitude (min/max)	Longitude (min/max)
Tibet-0	China	N90.98	E29.3
Ler-0	Poland	N50 / N55	E14.5 / E23
No-0	Germany	N51 / N51	E13 / E14
Wil-2	Russia	N55 / N55	E25 / E25
Kn-0	Lithuania	N54 / N55	E23 / E24
Ws-0	Russia	N52 / N53	E30 / E30
Ct-1	Italy	N37 / N38	E15 / E15
Rsch-4	Russia	N56 / N57	E34 / E34
Tsu-0	Japan	N34 / N35	E136 / E129
Hi-0	Netherlands	N52 / N53	E5 / E6
Oy-0	Norway	N60	E6
Po-0	Germany	N50 / N51	E7 / E7
Edi-0	United Kingdom	N56 / N56	E3 / E3
Wu-0	Germany	/	/
Col-0	USA	N38 / N39	W92 / W93
Zu-0	Switzerland	N47 / N49	E8 / E9

Sf-2	Spain	N41 / N42	E3 / E3
Mt-0	Libya	N33 / N33	E23 / E23
Bur-0	Ireland	N52 / N55	W6 / W10
Can-0	Spain	N28 / N28	W15 / W15
Cvi-0	Cabo Verde	N15	W24
IP-Vim-0	Spain	N41	W7
Don-0	Spain	N36	W7
IP-Vis-0	Spain	N39	W7
IP-Her-12	Spain	N39	W6
IP-Con-0	Spain	N37	W6
IP-Gra-0	Spain	N36	W6
IP-Iso-4	Spain	N43	W6
IP-Moj-0	Spain	N36	W6
IP-Pun-0	Spain	N40	W5
IP-Cem-0	Spain	N41	W5
IP-Fun-0	Spain	N40	W5
IP-Ven-0	Spain	N40	W5
IP-Nac-0	Spain	N40	W4
IP-Mar-1	Spain	N39	W4

Ped-0	Spain	N40	W4
IP-Cat-0	Spain	N40	W4
IP-Hum-2	Spain	N42	W4
IP-Sne-0	Spain	N37	W4
IP-Lso-0	Spain	N38	W4
IP-Cor-0	Spain	N40	W2
IP-Per-0	Spain	N37	W2
IP-Alm-0	Spain	N39	W1
Etna-2	Italy	N37	E14
Qar-8a	Lebanon	N34	E35
Kas-1	India	N35	E77
Kas-2	India	N35	E77
Altai-5	China	N47	E88

Table S5. Reads of the genome-wide resequencing of Tibet-0

All Reads	Mapped Reads	Mapped Ratio(%)	Unique mapped	Multiple mapped	Unmapped
109382230	100807134	0.92	77242012	23565122	8575096

Table S6. Genome-wide comparison between Tibet-0 and Col-0

Type	Reference (TAIR 10)	Tibet-0
genes	28775	28647
gene models	41671	41458
mRNA	35386	35265
CDS	35386	35237
pseudogenes	924	906
pseudogenic_transcript	926	908
transposable_element_gene	3903	3855
mRNA_TE_GENE	3911	3863
ncRNA	480	474
miRNA	180	179
rRNA	15	15
snoRNA	71	71
snRNA	13	13
tRNA	689	683
total	152330	151574

Table S7. Genome-wide SNPs of Tibet-0

	SNP	Intergenic	Upstream	5' UTR	CDS	Intron	3' UTR	Downstream	Synonymous	Non-synonymous
Chr1	296860	102986	35358	5686	59156	44394	9875	26365	30816	27770
Chr2	218482	107499	23839	3215	30148	24706	5379	16241	15840	14030
Chr3	247670	112898	26458	4098	40318	28335	6160	18900	20162	19680
Chr4	218243	97723	22232	3143	37284	28251	5817	15398	19782	17131
Chr5	279959	105586	32712	4575	52784	39795	8211	23600	27844	24514
ChrM	74	23	14	2	19	6	0	4	5	14
ChrC	59	1	12	0	21	4	0	5	11	10
Total	1261347	526716	140625	20719	219730	165491	35442	100513	114460	103149

Table S8. Genome-wide Indels of Tibet-0

	Indel	Intergenic	Upstream	5' UTR	CDS	Intron	3' UTR	Downstream	Non-frame-shift	Frame-shift
Chr1	38081	9590	7369	1770	1338	8364	2326	4492	726	575
Chr2	23546	7283	4470	903	790	4644	1374	2512	422	351
Chr3	27340	7761	5067	1245	1058	5333	1554	2998	541	487
Chr4	23912	6391	4447	977	907	5282	1433	2697	490	386
Chr5	33523	8291	6659	1450	1278	7233	2037	3957	718	537
ChrM	18	7	0	0	3	4	0	2	0	3
ChrC	1	0	0	0	0	0	0	1	0	0
Total	146421	39323	28012	6345	5374	30860	8724	16659	2897	2339

Table S9. SNPs and RNA-seq information of 19 *A. thaliana* sequences. (Gan et al. 2011)[1]

Accession	Nucleotide differences		RNA-seq support (SNPs)		
	SNPs	Ambiguous positions	Agree	Disagree	%concordance
Can-0	789,187	54,148	118,965	343	99.7125
Bur-0	673,965	51,113	101,981	248	99.7574
Ct-1	650,332	44,342	100,175	271	99.73
Edi-0	630,728	35,210	93,151	356	99.62
Hi-0	497,688	157,526	78,702	164	99.79
Kn-0	637,034	43,211	98,204	265	99.73
Ler-0	647,094	42,652	98,421	256	99.74
Mt-0	588,481	41,237	84,438	266	99.69
No-0	611,346	45,886	87,517	233	99.73
Oy-0	639,949	42,316	83,055	287	99.65
Po-0	446,422	267,439	57,604	688	98.82
Rsch-4	584,081	44,225	76,089	222	99.71
Sf-2	671,638	72,661	87,741	234	99.73
Tsu-0	615,062	43,030	85,738	243	99.72
Wil-2	661,673	48,184	89,111	316	99.65
Ws-0	652,654	47,181	84,478	207	99.76
Wu-0	592,611	47,585	82,864	227	99.73
Zu-0	631,624	44,391	92,757	219	99.76
Nonredundant	SNPs=3,071,117		SNPs with RNA-seq support = 503,825		

Materials and Methods:

Plant material and treatments

The seeds and leaves of Tibet-0 were collected in the wild forest of Gongga County (N90.98, E29.3, altitude: 4200), Tibet in 2013 when it had blossomed and borne fruit. The leaves were immediately dried with silica gel and stored in sealed bags for the following molecular identification. We then grew the Tibet-0 under a 16h light (22℃)/8 h dark (18℃) photoperiod regime and collected the flowers and leaves for further tests and sequencing.

We also used seeds of Col-0, Ler-0, Bur-0, Ws-0, Can-0 and Tibet-0 in the common garden experiments. There are 6 replicates of Can-0 and 13 replicates of each of the other accessions. Tibet-0 and five other *A. thaliana* ecotypes were grown at 22℃ under a 16h light/8 h dark photoperiod regime. The phenotypes of the five ecotypes from germination to fruit were recorded and compared.

Molecular Identification by gene barcoding

In order to prove Tibet-0 to be *A. thaliana*, we compared the nuclear internal transcribed spacer (ITS), four chloroplast genes (matK, rbcL, rpoB, rps16) and two chloroplast intergenic spacers (IGS) (trnL-trnF, trnT-trnL) between *A. thaliana* and A. lyrata (Online , Simon, Trajanoski et al. 2012). The total DNA was extracted by a TIANGEN Plant Genomic DNA Kit. We used the universal primers of each barcoding sequence in PCR amplification. The PCR reaction system contained 10ng DNA template, 2ul (10μM) forward-reverse primers, 25 μl 2x Taq PCR Master Mix (Tiangen Biotech). The PCR procedure consisted of one cycle of 94℃ 5min, 32 cycles of 94℃ 1min, Ta℃ (Supplementary Table 1) 1 min, and 72℃ 1.5 min, and one cycle of 72℃ 7min and 10℃10min. The PCR products were sequenced using an ABI 3730 automated DNA sequencer (Applied Biosystems, Foster City, California, USA). We then compared all the 7 sequences of the Tibetan sample with the homologous sequence of Col-0 and Arabidopsis lyrata downloaded from NCBI by alignment using CLUSTALX (Thompson, Gibson et al. 1997). We also tested coverage and identity of those genes in the Tibetan sample, Col-0, and A. lyrata using the Blast tool of NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Polyploidy determination

To determine the polyploidy of the new *Arabidopsis thaliana*, the seeds of *A. thaliana* ecotypes Columbia and Tibetan were grown on filter paper with distilled water at 22℃ with 16hr light/8hr dark cycles. The seedlings (4-5 days old, about 1cm long) were transferred on filter paper with 0.002M hydroxyquinoline solution for 2 hours at room temperature and for 2 hours at 4℃ in the dark. Roots of seedlings were

submerged in hydroxyquinoline solution. The root tips were isolated and fixed in ethanol/glacial acetic acid (3:1) for one hour at room temperature and overnight at 4℃. The root tissue was then incubated in carbolfuchsin solution for 2-3 hours at room temperature. The root tips were transferred onto a slide, and covered by a cover slip. Most of the residual staining solution was removed by once strongly pressing a flat hand on the slide (with cover slip) wrapped with filter paper. The preparations were viewed in a light microscope.

The pollen of Tibet-0 was also observed. Seeds of wild Tibet-0 were grown under a 16h light (22℃)/8 h dark (18℃) photoperiod regime until blossoming occurred. The flower bud of Tibet-0 (about 0.8–1mm) was collected at 10 am, fixed with Carony's fixative (ethanol: acetic acid =3:1) for 24 h and then stored in 70% alcohol at 4℃. Following cleaning with distilled water, the pollen mother cell was then squeezed out from anther to a slide, with one drop of mixed enzyme containing 0.3% cellulase, 0.3% pectinase, and 0.5% snailase at 37℃ for 10 min. Using filter paper to remove the excess enzyme solution, we then added a drop of dyeing liquid (1.5ug/ml DAPI, H-1500.vector) and covered the sample with coverslips, pressing it slightly. After staining in dark for 10 min, we observed and photographed the sample under a fluorescence microscope.

Genome-wide resequencing

The total DNA was extracted by a TIANGEN Plant Genomic DNA Kit, and genome-wide sequenced with a mean coverage of 40x the reference genomes Col-0, TAIR10 which are available at ftp://ftp.arabidopsis.org/Genes/TAIR10_genome_release/, by an Illumina HiSeq 2000 system.

Sequence preparation and orthologus identification

The gene annotation of the reference genome of *A. thaliana* was downloaded from https://www.arabidopsis.org. The CDs sequence of other 47 *A. thaliana* ecotypes including Col-0, Bur-0, Ct-1, Edi-0, Hi-0, Kn-0, Ler-0, Mt-0, No-0, Po-0, Oy-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0, Zu-0, Kas-1, Kas-2, Altai-5 and 26 relicts including Qar-8a, Etna-2, IP-Alm-0, IP-Cor-0, IP-Sne-0, IP-Con-0, Don-0, IP-Per-0, IP-Gra-0, IP-Moj-0, IP-Her-12, IP-Lso-0, IP-Cat-0, IP-Mar-1, IP-Vis-0, IP-Cem-0, IP-Vim-0, IP-Hum-2, Ped-0, IP-Nac-0, IP-Pun-0, IP-Ven-0, IP-Fun-0, Can-0, Cvi-0, IP-Iso-4 (Genomes Consortium. Electronic address and Genomes 2016) were downloaded from http://1001genomes.org/. The A. lyrata is an outcrossing perennial relative of *A. thaliana*. The sequence of A. lyrata was downloaded from Phytozome v9.1 (http://www.phytozome.net/) as outgroup. We first concatenated the CDS sequence of each gene according to its direction. We then searched the

orthologous genes of the Tibet-0, 47 other *A. thaliana* ecotypes (26 relicts, Col-0, Bur-0, Can-0, Ct-1, Edi-0 Hi-0, Kn-0, Ler-0, Mt-0, No-0, Po-0, Oy-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0, and Zu-0) and A. lyrata by a bidirectional-best-hit method in BLAST, and obtained a total of 5741 single-copy orthologous genes. The genes of which less than half of the total CDS length were covered were then eliminated. Finally, we aligned 5611 orthologous genes in 48 *A. thaliana* ecotypes.

Nucleotide Diversity

The scaled mutation rate $\theta\omega = 4Ne\mu$ was estimated by using the proportion of segregation sites θs (Watterson 1975) and the average pairwise nucleotide diversity $\theta\pi$ (Tajima 1989), which derived to the Tajima's D to demonstrate the distribution of the segregating sites. Putative genes were included in the analyses only if resequence data covered more than 50% of the putative coding sequences from all the accessions. As a result, a total of 5611 single copy orthologous genes of 48 *A. thaliana* ecotypes were applied in the calculation. Among which, 103 genes containing no segregating sites were eliminated in the Tajima's D deduction. Analyses were conducted of the tajima_d function provided from DendroPy Python library (Sukumaran and Holder 2010), available R codes, or custom R or Python scripts.

Phylogenetic analysis

We aligned each of the 5611 orthologues genes of 48 *A. thaliana* ecotypes and concatenated them to long orthologous alignments to construct the maximum-likelihood (ML) tree, using RAxML v8.0 (Stamatakis 2014) . The ML tree applied a partition model to take account of the different evolution rates of the first, second and third position of the codon. GTRGAMMA model was used for nucleotide substitution model. The confidence of the tree topology was evaluated by the bootstrap method with 100 replications. The time of the most recent common ancestor (tMRCA) was calculated by the Bayesian coalescent method with the BP&P program (Rannala and Yang 2003, Yang and Rannala 2010). We then verified the accuracy of the phylogenetic tree by calculating the time of the most recent common ancestor (tMRCA) by the coalescent method (Song, Liu et al. 2012, Nakagome, Nakajima et al. 2013). Single-copy orthologous genes in 47 species and longer than 1000bp were extracted, and the tMRCA48 value of each gene was obtained using Bpp software (Yang 2015). We then removed each ecotype to form 48 new datasets and calculated the tMRCA47 values. For each dataset, the difference between each tMRCA47 and tMRCA48 was detected by using one tail distribution pairwise t-test.

Divergence time estimation

We used two steps to estimate the divergence time of Tibet-0 and other ecotypes. First, we downloaded the published genome sequence of 23 species including 21 flowering, 1 fern (Selaginella moellendorffii) and 1 moss (Physcomitrella patens) from Phytozome v9.1 (http://www.phytozome.net/) (Goodstein, Shu et al. 2012) before March. 24th, 2014, and downloaded the whole genome sequence of Norway spruce (Picea abies) from Spruce Genome Project (ftp://congenie.org/congenie/) (Supplementary Table 2). After bidirectional blast between the sequence of these 24 species and the reference genome of Col-0, and removing paralogue contamination, we obtained 334 single-copy orthologous genes shared by 22 species or more. We concatenated the aligned 334 genes in each species and used '?' to fill the deletion. Based on the nucleotide and amino acid sequence super matrix, the phylogenetic ML tree was then constructed using RAxML v8.0 (Stamatakis 2014). The nucleotide ML tree applied a partition model to normalize the difference between evolution rates of the first, second and third position of the codon, and the model for nucleotide substitution rate was GTR+I+G, while the amino acid ML tree was constructed using a GAMMA+LG4XF amino acid substitution rate model. Both trees were bootstrapped for 1000 times. The divergence time was estimated by using MCMCTREE package in PAML software (Yang and Rannala 2006, Rannala and Yang 2007). The model for nucleotide sequence is a GTR substitution rate model, while the model for amino acid sequence is a F84 substitution rate model. We used the relative rate test and the likelihood ratio test to detect each branch of the phylogenetic tree. In this study, 7 fossil calibrations were used to correct the divergence time of each branch (Supplementary Table 3), thus obtain the divergence time between *A. thaliana* and A. lyrata.

Secondly, we estimated the time of common ancestor between Tibet-0 and the other 47 ecotypes using BP&P software based on Bayesian coalescent models (Yang 2002, Rannala and Yang 2003, Burgess and Yang 2008, Yang and Rannala 2010). By using ratio of the tau of Tibet-0 minus other thaliana accessions and the tau of the *A. thaliana* minus A. lyrata, we obtained the ratio of the divergence time between the *A. thaliana* and the A. lyrata and the divergence time between Tibet-0 and other *A. thaliana* accessions. Since we already obtained the divergence time between *A. thaliana* and A. lyrata, the divergence time between the Tibet-0 and other accessions could be calculated by multiplying the ratio and the divergence time between *A. thaliana* and A. lyrata.

Positive selection detection

The genomic DNA reads of other 19 *A. thaliana* ecotypes including Col-0, Bur-0, Can-0, Ct-1, Edi-0, Hi-0, Kn-0, Ler-0, Mt-0, No-0, Po-0, Oy-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0, and Zu-0 were downloaded from

http://mus.well.ox.ac.uk/19genomes/ (Gan, Stegle et al. 2011), which are also deposited in the European Nucleotide Archive (www.ebi.ac.uk/ena/) under accession number ERP000565(Gan, Stegle et al. 2011). According to the difference between Tibet-0 and other ecotypes shown in the common garden experiments, we investigated the potential adaptive evolution on the relative genes by using the CODEML package in the PAML software (Yang 2007). We applied the "branch-site" model that divided the branches of phylogenetic trees into foreground and background (Yang 1998). In order to detect if the genes in Tibet-0 have experienced positive selection, we regarded Tibet-0 as the foreground branch, with the other 19 ecotypes the background branches. If an enzyme has experienced positive selection, we searched for its protein structure in the PDB database (http://www.rcsb.org/pdb/home/home.do), and observed whether its structure was affected by the selection by using Pymol (http://www.pymol.org).

Supplementary Reference

Burgess, R. and Z. Yang (2008). "Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors." Mol Biol Evol 25(9): 1979-1994.

Gan, X., O. Stegle, J. Behr, J. G. Steffen, P. Drewe, K. L. Hildebrand, R. Lyngsoe, S. J. Schultheiss, E. J. Osborne, V. T. Sreedharan, A. Kahles, R. Bohnert, G. Jean, P. Derwent, P. Kersey, E. J. Belfield, N. P. Harberd, E. Kemen, C. Toomajian, P. X. Kover, R. M. Clark, G. Ratsch and R. Mott (2011). "Multiple reference genomes and transcriptomes for Arabidopsis thaliana." Nature 477(7365): 419-423.

Genomes Consortium. Electronic address, m. n. g. o. a. a. and C. Genomes (2016). "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana." Cell 166(2): 481-491.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam and D. S. Rokhsar (2012). "Phytozome: a comparative platform for green plant genomics." Nucleic Acids Res 40(Database issue): D1178-1186.

Nakagome, S., Y. Nakajima and S. Mano (2013). "Biogeography revealed by mariner-like transposable element sequences via a Bayesian coalescent approach." J Mol Evol 77(3): 64-69.

Online, S. s. m. o. S.

Rannala, B. and Z. Yang (2003). "Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci." Genetics 164(4): 1645-1656.

Rannala, B. and Z. Yang (2007). "Inferring speciation times under an episodic molecular clock." Syst Biol 56(3): 453-466.

Simon, U. K., S. Trajanoski, T. Kroneis, P. Sedlmayr, C. Guelly and H. Guttenberger (2012). "Accession-specific haplotypes of the internal transcribed spacer region in

Arabidopsis thaliana--a means for barcoding populations." Mol Biol Evol 29(9): 2231-2239.
Song, S., L. Liu, S. V. Edwards and S. Wu (2012). "Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model." Proc Natl Acad Sci U S A 109(37): 14942-14947.
Stamatakis, A. (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." Bioinformatics 30(9): 1312-1313.
Sukumaran, J. and M. T. Holder (2010). "DendroPy: a Python library for phylogenetic computing." Bioinformatics 26(12): 1569-1571.
Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics 123(3): 585-595.
Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins (1997). "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." Nucleic Acids Res 25(24): 4876-4882.
Watterson, G. A. (1975). "On the number of segregating sites in genetical models without recombination." Theor Popul Biol 7(2): 256-276.
Yang, Z. (1998). "On the best evolutionary rate for phylogenetic analysis." Syst Biol 47(1): 125-133.
Yang, Z. (2002). "Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci." Genetics 162(4): 1811-1823.
Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol 24(8): 1586-1591.
Yang, Z. (2015). "A tutorial of BPP for species tree estimation and species delimitation " Current Zoology 61: 854-865.
Yang, Z. and B. Rannala (2006). "Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds." Mol Biol Evol 23(1): 212-226.
Yang, Z. and B. Rannala (2010). "Bayesian species delimitation using multilocus sequence data." Proc Natl Acad Sci U S A 107(20): 9264-9269.